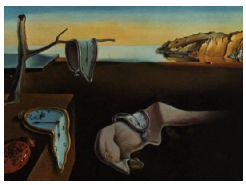

Journal of Statistical Software
January 2010, Volume 32, Code Snippet 2. http://www.jstatsoft.org/

R-Squared Measures for Two-Level Hierarchical Linear Models Using SAS

Anthony Recchia
University of Illinois

Abstract

The hierarchical linear model (HLM) is the primary tool of multilevel analysis, a set of techniques for examining data with nested sources of variability. The concept of R^2 from classical multiple regression analysis cannot be applied directly to HLMs without certain undesirable results. However, multilevel analogues have been formulated. The goal here is to demonstrate a SAS macro that will calculate estimates of these quantities for a two-level HLM that has been fit with SAS's linear mixed modeling procedure, `PROC MIXED`.

Keywords: hierarchical linear model, `PROC MIXED`, R-squared, SAS.

1. Introduction

1.1. Multilevel analysis

Multilevel analysis is a set of statistical techniques for examining data with sources of variability that are nested within one another. Data that possess such a structure arise frequently in practice. The simplest and most common form is two-level data, in which "level-1 units", or "individuals", are nested within "level-2 units", or "groups". Some typical examples of this scheme include students nested within classes, employees nested within firms, and measurements nested within subjects (in the case of longitudinal data).

Snijders and Bosker (1999) give two reasons why this type of data tends to appear. First, it is often the case that there are relationships between level-1 and level-2 units that are worthy of research. For instance, in the case of students nested within classes, two-level data on students' achievement could reveal how it is affected by differences among their teachers. Secondly, two-stage sampling is often more practical and efficient than simple random sampling. That is, in

our previous example, randomly sampling individual students from the population would cost far more in time and money than would randomly sampling students from a much smaller random sample of classes.

1.2. Hierarchical linear models

Modeling two-level data requires a more general technique than classical multiple linear regression analysis. This is because multiple linear regression includes an underlying assumption of residuals that are independent and identically distributed. Such an assumption could easily be inappropriate in the two-level case since there is likely to be dependence among the individuals that belong to a given group. For instance, it would be difficult to imagine that the academic achievements of students in the same class were not somehow related to one another.

The primary tool of multilevel analysis is the hierarchical linear model (HLM), which is formulated as follows. Suppose that our data consist of N groups with n_j individuals in group j. Let Y be the response variable and let $X_1, \ldots, X_p$ and $Z_1, \ldots, Z_q$ be not necessarily disjoint lists of explanatory variables. These lists contain the fixed- and random-effects regressors, respectively. They are treated differently by the model in that the regression coefficients for the former list of regressors will be constants while those for the latter list will be random variables. Nevertheless, the members of both lists are treated as random variables for the purpose of defining R-squared measures.

In equation form, the model for the response of the i-th individual in the j-th group is

$$Y_{ij} = \underbrace{\sum_{k=0}^{p} \beta_k X_{kij}}_{\text{Expectation}} + \underbrace{\sum_{h=0}^{q} U_{hj} Z_{hij} + E_{ij}}_{\text{Residual}} .$$

Here we set $X_{0ij} = Z_{0ij} = 1$ for all i and j to account for the fixed and random intercepts, if applicable.

It is allowed that some of the fixed-effects regressors $X_1, \ldots, X_p$ take different values for different individuals while others take the same value for every member of the same group. That is, one may make use of variables that describe the individual ("level-1 variables") as well as variables that describe the group to which the individual belongs ("level-2 variables"). The response variable Y and the random-effects regressors $Z_1, \ldots, Z_q$, however, must always be level-1 variables.

Additionally, we make the following assumptions:

- The fixed-effects regressor vectors $\mathbf{X}_{ij} = (X_{1ij}, \ldots, X_{pij})$ are identically distributed.
- The random effect vectors $\mathbf{U}_j = (U_{0j}, \ldots, U_{qj})^\top$ are independent and identically distributed as $N(\mathbf{0}, \mathbf{T})$.
- The random errors E_{ij} are independent and identically distributed as $N(0, \sigma^2)$.
- The regressors X_{kij} and Z_{hij} are independent of the random effects U_{hj} and errors E_{ij}.

- The random effects U_{hj} are independent of the random errors E_{ij}.

These assumptions express, among other things, the idea that the responses of individuals in the same group may not be independent of one another even though there is independence among the groups. Whereas the residual of a multiple linear regression model would only include the E_{ij} term, it is this term along with the random effects summation that account for the residual of a HLM. The residuals of every member of group j will depend upon the same vector $\mathbf{U}_j$ of random effects, so there is dependence among these residuals. This is meant to separate the variation accounted for by group membership from variation due to other sources.

2. R-squared statistics for HLMs

2.1. Complications

As in all statistical analyses, it is desirable to have statistics that will help a researcher to assess how well his/her HLM is performing. Multiple linear regression analysis has R^2, the proportional reduction in the single variance component of the model. Although it might be tempting to apply this idea to each of the variance components in a HLM separately, Snijders and Bosker (1994) warn that doing so can produce undesirable results.

The complications arise from the fact that variation in the response variable of a two-level HLM is assumed to come from multiple sources, namely the two levels underlying the data. Descriptions of the variability at each level require both the random effects covariance matrix $\mathbf{T}$ and the error variance σ^2, but estimates of these variance components do not necessarily behave as one might expect. Indeed, the addition of an explanatory variable to a HLM can simultaneously increase some of the variance components and decrease others. This means that examining the individual components of variance separately by way of a traditional R^2 can lead to surprising outcomes like negative values or values that decrease when a new regressor is added to the model.

Instead, Snijders and Bosker (1994) suggest separate examinations of the levels of variance. They show that the population values of the resulting measures possess the appealing properties that they are always nonnegative and that additional explanatory variables will never cause them to decrease, assuming that the fixed-effects portion of the model is specified correctly. However, as of yet, no software directly computes these multilevel analogues of the standard R^2. Therefore, the goal here is the demonstration of a SAS (SAS Institute Inc. 2003) macro that will calculate estimates of these quantities for a two-level HLM that has been fit with SAS's linear mixed modeling procedure, `PROC MIXED`.

2.2. Formulas

Although R^2 is usually defined in the multiple linear regression setting as "explained proportion of variance", the approach of Snijders and Bosker (1994) employs the equivalent definition of "proportional reduction in mean squared prediction error". One wishes to predict an individual's value Y_{ij} of Y at level 1 and a group's mean value $\overline{Y}_{\cdot j} = \frac{1}{n_j}\sum_{i=1}^{n_j} Y_{ij}$ at level 2, so it is possible to use this alternative definition to derive separate measures of proportional reduction in mean squared prediction error for each level.

Suppose that we have a HLM as described above, and make the following definitions:

- $\boldsymbol{\mu}_Z$ is the expectation vector of the random-effects regressors.
- $\boldsymbol{\Sigma}_Z^B$ is the between-group covariance matrix of the random-effects regressors.
- $\boldsymbol{\Sigma}_Z^W$ is the within-group covariance matrix of the random-effects regressors.
- $\tilde{\tau}^2$ and $\tilde{\sigma}^2$ are the variance components from the so-called "empty model". This is the model that contains a fixed and random intercept but no regressors. In equation form, it is given by $Y_{ij} = \tilde{\beta}_0 + \tilde{U}_{0j} + \tilde{E}_{ij}$ where the $\tilde{U}_{0j}$ are independent and identically distributed as $N(0, \tilde{\tau}^2)$, the $\tilde{E}_{ij}$ are independent and identically distributed as $N(0, \tilde{\sigma}^2)$, and all pairs of $\tilde{U}_{0j}$ and $\tilde{E}_{ij}$ are independent.
- n is a representative group size.

The empty model described above is essentially the simplest model that still makes use of group membership information. Therefore, it is utilized as the baseline from which the proportional reduction in mean squared prediction error is computed.

First, considering level 1, Snijders and Bosker (1994) show that the mean squared prediction error from predicting Y_{ij} using its expectation under the model of interest is

$$\begin{aligned} \mathsf{VAR}\left(Y_{ij} - \sum_{k=0}^{p} \beta_k X_{kij}\right) &= \mathsf{VAR}\left(\sum_{h=0}^{q} U_{hj} Z_{hij} + E_{ij}\right) \\ &= \boldsymbol{\mu}_Z^\top \mathbf{T} \boldsymbol{\mu}_Z + \text{trace}\left[\mathbf{T}\left(\boldsymbol{\Sigma}_Z^B + \boldsymbol{\Sigma}_Z^W\right)\right] + \sigma^2 \end{aligned}$$

while that from using its expectation under the empty model is

$$\begin{aligned} \mathsf{VAR}\left(Y_{ij} - \tilde{\beta}_0\right) &= \mathsf{VAR}\left(\tilde{U}_{0j} + \tilde{E}_{ij}\right) \\ &= \tilde{\tau}^2 + \tilde{\sigma}^2. \end{aligned}$$

Hence, the proportional reduction in mean squared prediction error at level 1, R_1^2, is given by

$$R_1^2 = 1 - \frac{\boldsymbol{\mu}_Z^\top \mathbf{T} \boldsymbol{\mu}_Z + \text{trace}\left[\mathbf{T}\left(\boldsymbol{\Sigma}_Z^B + \boldsymbol{\Sigma}_Z^W\right)\right] + \sigma^2}{\tilde{\tau}^2 + \tilde{\sigma}^2}.$$

Next, considering level 2, Snijders and Bosker (1994) show that the mean squared prediction error from predicting $\overline{Y}_{\cdot j}$ using its expectation under the model of interest is

$$\begin{aligned} \mathsf{VAR}\left(\overline{Y}_{\cdot j} - \sum_{k=0}^{p} \beta_k \overline{X}_{k\cdot j}\right) &= \mathsf{VAR}\left(\sum_{h=0}^{q} U_{hj} \overline{Z}_{h\cdot j} + \overline{E}_{\cdot j}\right) \\ &= \boldsymbol{\mu}_Z^\top \mathbf{T} \boldsymbol{\mu}_Z + \text{trace}\left[\mathbf{T}\left(\boldsymbol{\Sigma}_Z^B + \frac{1}{n_j}\boldsymbol{\Sigma}_Z^W\right)\right] + \frac{\sigma^2}{n_j} \end{aligned}$$

while that from using its expectation under the empty model is

$$\begin{aligned} \mathsf{VAR}\left(\overline{Y}_{\cdot j} - \tilde{\beta}_0\right) &= \mathsf{VAR}\left(\tilde{U}_{0j} + \overline{\tilde{E}}_{\cdot j}\right) \\ &= \tilde{\tau}^2 + \frac{\tilde{\sigma}^2}{n_j}. \end{aligned}$$

Since the group sizes need not all be the same, the representative group size n must be used in their place in the above two expressions. Making these substitutions yields R_2^2, the proportional reduction in mean squared prediction error at level 2:

$$R_2^2 = 1 - \frac{\boldsymbol{\mu}_Z^\top \mathbf{T} \boldsymbol{\mu}_Z + \operatorname{trace}\left[\mathbf{T}\left(\boldsymbol{\Sigma}_Z^B + \frac{1}{n}\boldsymbol{\Sigma}_Z^W\right)\right] + \frac{\sigma^2}{n}}{\tilde{\tau}^2 + \frac{\tilde{\sigma}^2}{n}}.$$

Estimates of the parts of the formulas for R_1^2 and R_2^2 are relatively easy to obtain. Those of $\mathbf{T}$, σ^2, $\tilde{\tau}^2$, and $\tilde{\sigma}^2$, come directly from the modeling procedure. The vector $\boldsymbol{\mu}_Z$ is estimated by the vector of sample means of the variables $Z_0, \ldots, Z_q$. As recommended by Snijders and Bosker (1994), the representative group size is taken to be either a user-supplied value or the harmonic mean of the group sizes, namely

$$\frac{N}{\sum_{j=1}^{N} \frac{1}{n_j}}.$$

To obtain an estimate of $\boldsymbol{\Sigma}_Z^B$, we replace the observed values z_{hij} of each random-effects regressor Z_h with the corresponding group mean $\overline{z}_{h\cdot j}$ and compute the sample covariance matrix. Finally, for an estimate of $\boldsymbol{\Sigma}_Z^W$, we replace the observed values z_{hij} with the group-centered versions $z_{hij} - \overline{z}_{h\cdot j}$ and compute the sample covariance matrix.

3. Computation using the macro HLMRSQ

3.1. Implementation

The SAS macro for computing the R-squared measures is called `HLMRSQ` and is included with this article in the file `hlmrsq.sas`. It is meant to take as input several data sets that `PROC MIXED` can produce while fitting a model, an optional `WHERE` expression for subsetting the modeling data, and an optional value for the representative group size n. It then extracts the necessary information about the model, uses this information to compute estimates of the pieces of the formulas discussed earlier, calculates the explained proportion of variance estimates for both levels, and displays the results.

This setup means that a modeler would follow the basic procedure outlined below which is elaborated upon in the next section.

1. Write the `PROC MIXED` step for fitting the desired model as usual.
2. Modify this step so that it will produce the output that the macro requires as input.
3. Run the modified `PROC MIXED` step.
4. Invoke the macro.

More specific information about the arguments to the macro is given in Table 1.

The macro is able to handle several of the options and statements that are available to modelers in `PROC MIXED`. For instance, the macro will automatically use the same estimation

Argument	Description	Usage
CovParms	The name given to the CovParms= data set in the ODS OUTPUT statement	Required
GMatrix	The name given to the G= data set in the ODS OUTPUT statement	Required
ModelInfo	The name given to the ModelInfo= data set in the ODS OUTPUT statement	Required
SolutionF	The name given to the SolutionF= data set in the ODS OUTPUT statement	Required
WhereExpr	The expression used in the WHERE statement to subset the modeling data	Optional; should be enclosed in the %nrstr() function
RepSize	The value for the representative group size	Optional; defaults to the harmonic mean of the group sizes if omitted

Table 1: Arguments for the macro HLMRSQ.

method for fitting the empty model that the modeler used for the original model. It will also accommodate any of the shorthand specifications of effects that can be used in either the MODEL or RANDOM statements of PROC MIXED. Any variables that are designated as classification variables in the CLASS statement will also be treated appropriately by the macro.

The macro will also determine if a BY statement was used in the original PROC MIXED step in order to simultaneously fit multiple models. This situation could arise if, for example, the modeler had fit the same two-level model to each third-level unit of a three-level data set. In this case, the macro will calculate and output the two R-squared estimates for each subset determined by the distinct combinations of levels of the BY variables.

Beyond that, the macro will determine if the input model was fit off of weighted data with the WEIGHT statement and will adjust for the weight variable in its computations. Finally, as mentioned earlier, the macro can account for any subsetting of the modeling data that was accomplished via a WHERE statement if the modeler supplies the WHERE expression as an argument to the macro.

3.2. Example

We will work through the use of the macro with an example. For the sake of reproducibility, a two-level HLM was simulated in SAS using code that is provided with this article in the file v32c02.sas. The resulting data set is called hlmdata, and it contains the following variables: the response variable Yij, the grouping variable Groupnumber that identifies the group to which each individual belongs, a continuous level-2 explanatory variable X1ij, continuous level-1 explanatory variables X2ij and X3ij, and a binary level-1 explanatory variable X4ij.

The first step in using the macro is constructing the PROC MIXED step as usual. This will normally consist of a specification of the fixed- and random-effects regressors and the grouping variable as well as any additional options. If, in our example, one were interested in fitting a two-level HLM for Yij with fixed coefficients for each of the explanatory variables except for X3ij, a random intercept, and random coefficients for X2ij and X4ij, then SAS code which would fit such a model is below.

```
proc mixed data=hlmdata;
    class groupnumber;
    model Yij = X1ij X2ij X4ij;
    random intercept X2ij X4ij / subject=groupnumber;
run;
```

Note that in fitting this model in SAS, we are implicitly assuming that $Z_1 = X_2$ and $Z_2 = X_4$ in our HLM definition.

The second step is to modify the original `PROC MIXED` step so that it will to produce the output data sets that will be used as input to the macro. These modifications will include an extra statement as well as a few extra options for the statements that were already there.

```
proc mixed data=hlmdata namelen=200;
    class groupnumber;
    model Yij = X1ij X2ij X4ij / s;
    random intercept X2ij X4ij / subject=groupnumber g;
    ods output CovParms=cov G=gmat ModelInfo=mod SolutionF=solf;
run;
```

The modifications are as follows:

- The option `NAMELEN=200` has been added to the `PROC MIXED` statement to ensure that the output data sets contain the full names of all effects.
- The option `S` has been added to the `MODEL` statement after a forward slash so that the output data set containing the fixed-effects solutions will be produced.
- The option `G` has been added to the `RANDOM` statement so that the output data set containing the estimated random-effects covariance matrix will be produced.
- The `ODS OUTPUT` statement has been added to name the output data sets that will be used as input to the macro.

The modified code can then be submitted in order to fit the desired model. If the fit is successful, then the final step is to invoke the macro itself, supplying to it the names of the four data sets that were produced by the `PROC MIXED` step. In our example, this would be accomplished with the following macro call in SAS.

```
%hlmrsq(CovParms=cov, GMatrix=gmat, ModelInfo=mod, SolutionF=solf);
```

Note that since the entire modeling data set was used to fit the model and since we are not choosing our own value for the representative group size, the `WhereExpr` and `RepSize` arguments have been omitted.

The output of the macro contains the representative group size that was used in the computations and the R-squared estimates for both levels. It is reproduced in Table 2.

Explained proportion of variance		
Rep size	Level 1	Level 2
114.83	0.221825	0.524735

Table 2: R-squared measures for the example model.

4. Interpretation

Interpretation of the numerical values of these R-squared statistics is straightforward given their definitions as explained proportions of variance for the two levels of the data. As far as typical behavior of these measures is concerned, Snijders and Bosker (1994) show that the addition of a level-2 explanatory variable – which, by definition, has no within-group variability – will leave the estimate of σ^2 unchanged but decrease the common variance of the random effects U_{0j} that are associated with the intercept of the model. This will leave the estimate of R_1^2 the same and increase the estimate of R_2^2. They also show that the addition of a purely level-1 explanatory variable – that is, a variable with no between-group variability – will decrease the estimate of σ^2 but actually slightly increase the variance of the U_{0j}. This will increase the estimate of R_1^2 but decrease the estimate of R_2^2.

These R-squared statistics are also useful for diagnostic purposes. As stated earlier, when the fixed-effects portion of the model is specified correctly, the population parameters R_1^2 and R_2^2 are always nonnegative and will not decrease when an explanatory variable is added to the model. However, Snijders and Bosker (1994) note that the estimators $\hat{R}_1^2$ and $\hat{R}_2^2$ of these quantities do not have these properties. Therefore, if negative values are obtained or if the estimates decrease when a regressor is added in the course of fitting candidate models, then the cause must either be random chance or a misspecification of the fixed effects. Snijders and Bosker (1999) recommend that the modeler suspect misspecification if either of the R-squared estimates decreases by 0.05 or more upon the addition of an explanatory variable.

One important type of misspecification ought to be considered when the R-squared measures display anomalous behavior. It is sometimes the case in multilevel modeling that the within-groups relationship between a response variable and a level-1 fixed-effects regressor will have a different direction from the between-groups relationship. Returning once again to the example of students nested within classes, suppose that the response variable is a measure of academic achievement and the regressor is a measure of perceived level of support from the teacher. It could be the case that at the classroom level, greater perceived support results in higher academic achievement, while at the student level, teachers tend to give more support to lower performing students than to higher performing students. To include only the raw version of such an explanatory variable in a HLM is to make an implicit and incorrect assumption of equal within- and between-group regression coefficients which will tend to increase mean squared prediction error at the group level and, in so doing, decrease the estimate of R_2^2, perhaps to the point where it becomes negative. One way to correct this problem is to replace the raw variable with separate fixed-effects regressors for its group-centered version and its group means, thus allowing their estimated coefficients to take values with opposite signs.

5. Conclusion

Statistics like these explained proportion of variance measures are not, of course, the ultimate factor that a modeler will use to identify his/her final model. However, they are useful as diagnostic tools and as one additional set of criteria among many which researchers can use to attempt to rank order what is often a large number of candidate models. And while the fact that software does not currently compute these two particular statistics does not mean that a modeler could not calculate them on his/her own, there is a certain amount of reliability and convenience in having access to a macro like the one that we have been working with here.

In addition to its calculation of the two R-squared measures, this macro may also be useful for providing a framework in which valuable model information can be extracted automatically from a run of `PROC MIXED` for use by other SAS procedures. The researcher can utilize this capability to perform analyses that are not currently available in SAS and to customize the output to his/her needs.

Acknowledgments

I wish to thank Carolyn J. Anderson of the Department of Educational Psychology, University of Illinois, for her considerable help and encouragement and Maria E. Muyot of the Department of Statistics, University of Illinois, for her outstanding instruction in SAS programming.

References

SAS Institute Inc (2003). *The SAS System, Version 9.1*. Cary. URL http://www.sas.com/.

Snijders TAB, Bosker RJ (1994). "Modeled Variance in Two-Level Models." *Sociological Methods & Research*, **22**(3), 342–363.

Snijders TAB, Bosker RJ (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Sage Publications, London.

Affiliation:

Anthony Recchia
Department of Statistics
University of Illinois
725 S. Wright St.
Champaign, Illinois 61820, United States of America
E-mail: recchia@illinois.edu

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 32, Code Snippet 2 *Submitted:* 2009-07-02
January 2010 *Accepted:* 2009-12-17